WHAT IS CLAIMED IS:

1. A method for matching a reference document with a plurality of corpus documents, the method comprising:

deriving semantic content of the reference document according to a hierarchical arrangement of semantic types; and

for each corpus document,

deriving semantic content of the corpus document according to the hierarchical arrangement of semantic types; and

producing a matching score for the corpus document by determining a relatedness between the corpus document and the reference document from the derived semantic content of the corpus document and the derived semantic content of the reference document.

2. The method recited in claim 1 wherein deriving semantic content of the reference document and deriving semantic content of the corpus document comprises:

creating tokenized elements from a text stream;

tagging each tokenized element with a grammatical category label; and

creating a root form for each tokenized and tagged element.

3. The method recited in claim 2 wherein deriving semantic content of the reference document and deriving semantic content of the corpus document further comprises assigning a semantic type within the hierarchical arrangement of semantic types to the root form.

4. The method recited in claim 1 wherein producing the matching score comprises determining a distance within the hierarchical arrangement between a semantic type that defines semantic content of the reference document and a semantic type that defines semantic content of the corpus document.

5. The method recited in claim 4 wherein determining the distance comprises accounting for a qualia relationship between types in the hierarchical arrangement.

6. The method recited in claim 5 wherein the qualia relationship comprises a direct qualia relationship.

7. The method recited in claim 5 wherein the qualia relationship comprises an indirect qualia relationship.

8. The method recited in claim 5 wherein the qualia relationship comprises a telic relationship.

9. The method recited in claim 5 wherein the qualia relationship comprises an agentive relationship.

10. The method recited in claim 4 wherein producing the matching score further comprises accounting for whether the semantic type that defines semantic content of the reference document and the semantic type that defines semantic content of the corpus document are in a subsumption relationship.

11. The method recited in claim 4 wherein producing the matching score further comprises applying a filtering function to increase importance of a smaller distance relative to a larger distance.

12. The method recited in claim 11 wherein the filtering function comprises a Gaussian function.

13. The method recited in claim 11 wherein the filtering function comprises an exponential function.

14. The method recited in claim 11 wherein the filtering function comprises a rectangular function.

15. The method recited in claim 1 further comprising ranking the plurality of corpus documents in accordance with the matching score for each corpus document.

16. The method recited in claim 1 wherein the plurality of corpus documents is categorized according to a categorization scheme and the reference document comprises an uncategorized document, the method further comprising categorizing the uncategorized document according to the categorization scheme with the matching score.

17. The method recited in claim 16 wherein the categorization scheme comprises a hierarchical categorization scheme.

18. The method recited in claim 17 wherein the plurality of corpus documents is comprised by a larger set of documents within the hierarchical categorization scheme.

19. The method recited in claim 1 wherein the reference document comprises a user query.

20. The method recited in claim 19 wherein the plurality of corpus documents comprises a plurality of sponsor web pages, the method further comprising generating an output interest statement with semantic structures derived from at least one of the reference document and the corpus document having the highest matching score.

21. The method recited in claim 1 wherein the reference document and the plurality of corpus documents are comprised by a document set, the method further comprising:

determining the matching scores for a plurality of divisions of the document set into the reference document and the corpus documents;

combining the matching scores for each document pair comprised by the document set; and

clustering documents within the document set by setting a threshold for the combined matching scores.

22. A method for categorizing an uncategorized document within a categorization scheme, the method comprising:

deriving semantic content of the reference document according to a hierarchical arrangement of semantic types;

performing a comparison of the semantic content of the uncategorized document with semantic content of documents previously categorized according to the categorization scheme; and

determining a category for the uncategorized document from the comparison.

23. The method recited in claim 22 wherein the categorization scheme comprises a hierarchical categorization scheme.

24. The method recited in claim 23 wherein performing the comparison comprises, for each level of the hierarchical categorization scheme:

producing a matching score for each unexcluded document categorized at such level; and

excluding documents at a level subordinate to such level from the matching score.

25. The method recited in claim 22 wherein determining a category for the uncategorized document comprises determining a plurality of categories for the document.

26. The method recited in claim 22 wherein performing a comparison comprises producing a matching score for each of the plurality of documents previously categorized by determining a relatedness with the uncategorized document.

27. The method recited in claim 26 wherein producing the matching score comprises determining a distance within the hierarchical arrangement between a semantic type that defines content of the uncategorized document and a semantic type that defines semantic content of the previously categorized document.

28. The method recited in claim 27 wherein determining the distance comprises accounting for a qualia relationship between types in the hierarchical arrangement.

29. The method recited in claim 27 wherein producing the matching score further comprises accounting for whether the semantic type that defines semantic content of the uncategorized document and the semantic type that defines semantic content of the previously categorized document are in a subsumption relationship.

30. The method recited in claim 27 wherein producing the matching score further comprises applying a filtering function to increase importance of a smaller distance relative to a larger distance.

31. A system for matching a reference document with a plurality of corpus documents, the system comprising:

a database configured for storing a hierarchical arrangement of semantic types; and

an engine in communication with the database configured to derive semantic content of the reference document and of each corpus document according to the hierarchical arrangement; and

produce a matching score between the reference document and each corpus document from the derived semantic content.

32. The system recited in claim 31 wherein the engine is further configured to rank each corpus document according to its matching score.

33. The system recited in claim 31 wherein the engine is configured to produce the matching score by determining a distance within the hierarchical arrangement.

34. The system recited in claim 33 wherein determining the distance comprises accounting for a qualia relationship between types in the hierarchical arrangement.

35. The system recited in claim 33 wherein the matching score is filtered to increase the importance of a smaller distance relative to a larger distance.

36. The system recited in claim 31 wherein the engine is in communication with the internet.

37. A system for categorizing an uncategorized document within a categorization scheme, the system comprising:

a database configured for storing a categorization for each of a plurality of previously categorized documents and for storing a hierarchical arrangement of semantic types; and

an engine in communication with the database configured to

derive semantic content of the uncategorized document and of each of the plurality of previously categorized documents according to the hierarchical arrangement; and

compare the semantic content of the uncategorized document with the semantic content of each of the plurality of previously categorized documents to determine a category for the uncategorized document.

38. The system recited in claim 37 wherein the categorization scheme comprises a hierarchical categorization scheme.

39. The system recited in claim 37 wherein the engine is configured to compare the semantic content by producing a matching score between the uncategorized document and each of the plurality of previously categorized documents.

40. The system recited in claim 39 wherein the engine is configured to produce the matching score by determining a distance within the hierarchical arrangement.

41. The system recited in claim 40 wherein determining the distance comprises accounting for a qualia relationship between types in the hierarchical arrangement.

42. The system recited in claim 40 wherein the matching score is filtered to increase the importance of a smaller distance relative to a larger distance.

43. The system recited in claim 37 wherein the engine is in communication with the internet.